# Measuring Gender Bias in Natural Language Processing: Incorporating Gender-Neutral Linguistic Forms for Non-Binary Gender Identities in Abusive Speech Detection

**Nasim Sobhani**
SFI Centre for Research Training in Machine Learning
Technological University Dublin
nasim.x.sobhani@mytudublin.ie

**Kinshuk Sengupta**
Microsoft
Dublin, Ireland
kinshuk.sengupta@microsoft.com

**Sarah Jane Delany**
SFI Centre for Research Training in Machine Learning
Technological University Dublin
sarahjane.delany@tudublin.ie

## Abstract

Predictions from Machine Learning models can reflect bias in the data on which they are trained. Gender bias has been shown to be prevalent in Natural Language Processing models. The research into identifying and mitigating gender bias in these models predominantly considers gender as binary, male and female, neglecting the fluidity and continuity of gender as a variable.

In this paper, we present an approach to evaluate gender bias in a prediction task, which recognises the non-binary nature of gender. We gender-neutralise a random subset of existing real-world hate speech data. We extend the existing template approach for measuring gender bias to include test examples that are gender-neutral. Measuring the bias across a selection of hate speech datasets we show that the bias for the gender-neutral data is closer to that seen for test instances that identify as male than those that identify as female.

## 1 Introduction

Natural Language Processing (NLP) models and systems are developed by using text content created by humans and they may incorporate biases that exist in the data. These biases can then be reflected in the results produced by these models and systems when they are used in downstream applications (Dixon et al., 2018; Park et al., 2018). Additionally, word embeddings, which are representations of words and sentences generated from large amounts of natural language text, may also exhibit and even magnify certain features of the data, such as gender stereotypes (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2017).

An issue with existing research is that it considers gender as binary neglecting the fluidity and continuity of gender as a variable (Stanczak and Augenstein, 2021). Many of the data resources in NLP currently are inadequate for identifying gender bias as they often have a significant under-representation of female or non-binary instances. (Stanczak and Augenstein, 2021). There is a need to incorporate gender-neutral linguistic forms in datasets and algorithms to recognise the non-binary nature of gender. This impacts on algorithms too, such as language models which learn meaningless unstable representations for non-binary associated pronouns and terms (Dev et al., 2021).

In this paper, we present an approach to measure gender bias in a downstream task to identify abusive or hate speech that considers male, female, and gender-neutral gender identities. Due to the lack of datasets that include gender-neutral linguistic forms, we adjust existing real-world datasets by gender-neutralising a random subset of instances.

A challenge with measuring gender bias in natural language training data is the lack of gender identification in the data. One solution to this is to generate synthetic test data with a known gender identity using a template approach known as GBETs (Sun et al., 2019). Our approach has extended existing binary template definitions to include identity terms that reflect gender neutrality. We use a suite of measures presented by Borkan et al. (2019b) which are threshold agnostic to measure gender bias.

The downstream task we address is abusive and hate speech which involves language that is intended to be harmful and specifically targets individuals based on their affiliation with a particular group, such as their race, gender, sexuality, religion, or other protected characteristics (Röttger et al., 2021). Hate speech detection models exhibit

*Proceedings of Recent Advances in Natural Language Processing*, pages 1121–1131
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_119

gender biases towards certain identity terms due to factors such as an uneven distribution of identity terms in hate speech datasets and the excessive use of certain identity terms in hate speech sentences. For instance, some terms, like "women" and "feminism," can be frequently associated with sexist comments in benchmark datasets. These factors can lead to overfitting of the original hate speech detection model, which in turn may result in incorrect generalisations, such as linking the word "women" with a "hateful" label (Park et al., 2018; Mozafari et al., 2020).

We evaluate gender bias on three real-world hate speech datasets that have been adjusted to include data instances with a gender-neutral identity. The findings show that the bias for gender-neutral data is closer to that seen for data that is identified as male than data that is identified as female.

The contribution of this work lies in its recognition and exploration of the non-binary nature of gender in the context of measuring and addressing bias in NLP models and systems. While previous research has primarily focused on gender as a binary variable, this study goes beyond the traditional binary categorization and acknowledges the fluidity and continuity of gender identities. By incorporating gender-neutral linguistic forms in datasets we aim to promote gender inclusion and recognise the non-binary spectrum of gender. This approach allows for a more comprehensive understanding of gender bias in NLP and provides insights into the biases present in hate speech detection models.

## 2 Related Work

In supervised learning contexts, there is significant research that identifies and measures bias in downstream NLP tasks. Gender and racial biases (Kiritchenko and Mohammad, 2018), as well as biases against queer individuals (Ungless et al., 2023) and people with disabilities (Hutchinson et al., 2020) have been identified in sentiment analysis tasks. Gendered occupational stereotypes are reflected in errors made by co-reference resolution systems (Zhao et al., 2018; Rudinger et al., 2017) and occupational classification models (De-Arteaga et al., 2019).

A wide range of research into gender bias studies predominantly focuses on two genders, male and female, not recognising the experiences of individuals who identify as non-binary or gender non-conforming. This is a significant limitation of much of the existing research, as it fails to fully capture the diverse experiences and perspectives of individuals across the gender spectrum. Recent research has highlighted the importance of including non-binary identities in NLP studies. Studies focusing on neopronouns have shown that language models have difficulties processing them in various languages, including Swedish, Danish, and English (Brandl et al., 2022). Also, work by Cao and Daumé III (2021) proposes methods for improving gender inclusivity throughout the Machine Learning lifecycle, including data collection, model training, and evaluation. A road map toward the integration of inclusive language in translation, with a focus on machine translation tasks, has been discussed in work by Piergentili et al. (2023). This work focuses on gender-neutralisation strategies in the context of English-Italian translation.

Moreover, in order to improve support for individuals who identify as non-binary or gender non-conforming, enabling them to self-identify their preferred pronouns and interact with technology in a manner that aligns with their social identity, gender-neutral rewriting models have emerged (Sun et al., 2021; Vanmassenhove et al., 2021) in the text generation task. The purpose of a gender-neutral rewriter is to automatically identify the gendered language in a text and replace it with gender-neutral alternatives. In order to produce gender-neutral language, research by Sun et al. (2021); Vanmassenhove et al. (2021) in a relatively similar approach proposed a rule-based and neural approach to automatically rewrite text to be more gender-neutral. The system is designed to identify gender identity words such as "he/she" and replace them with "they". The goal is to promote inclusivity and reduce bias in language use by avoiding gender-specific language that may reinforce gender stereotypes or exclude individuals who do not identify with traditional gender roles.

### 2.1 Measuring Gender Bias

The primary method to measure gender bias in a downstream task is to measure performance differences across gender as the system's performance should not be influenced by gender. This requires a way to isolate gender in the test instances which are used to measure the system performance. While it is possible to isolate and identify gender for some types of training data, e.g. job applications in recruitment, for most textual corpora there is no obvious gender identification. Gender identification is typically done by generating synthetic test sets that contain test instances designed to isolate a particular group. This method is referred to as Gender Bias Evaluation Testsets (GBETs), as named by Sun et al. (2019), and has been used to evaluate bias in various NLP tasks.

GBETs have been categorised into three groups (Stanczak and Augenstein, 2021), template-based datasets, natural language-based datasets, and datasets generated for probing language models. The template approach involves creating sentence templates that include gender identification words that are relevant to the specific downstream task. From these templates, pairs of sentences are generated for each gender, and the performance of the NLP system is compared across the sentences with male and female gender identities, allowing for the measurement of gender bias in the dataset. This gender identity template approach has been used for various NLP tasks, including abusive language detection (Dixon et al., 2018; Park et al., 2018), sentiment analysis (Kiritchenko and Mohammad, 2018), and coreference resolution (Zhao et al., 2018; Rudinger et al., 2017).

Natural language-based GBET datasets use available natural language resources created in different ways, depending on the specific NLP task being evaluated. For instance, the GAP corpus (Webster et al., 2018) is a GBET used for coreference resolution and consists of ambiguous pronoun-name pairs that have been manually labeled by humans and sourced from Wikipedia. Similarly, (Emami et al., 2019) created a dataset for analysing gender bias in coreference resolution by scraping data from sources such as Wikipedia, OpenSubtitle, and Reddit comments.

More recently StereoSet (Nadeem et al., 2021) and CrowS-Pairs (Nangia et al., 2020) GBETs have been proposed to evaluate bias in language models. These GBETs are created and annotated by crowdsourcing to measure bias in different domains. Each example consists of a pair of stereotype and anti-stereotype sentences in the case of CrowS-pairs. However, StereoSet contains triplets of sentences with each instance corresponding to a stereotypical, anti-stereotypical, or meaningless association. An additional study presents a large GBET dataset called HOLISTICBIAS for measuring bias. This dataset is assembled by using a set of demographic descriptor terms in a set of bias measurement templates and can be used to test bias in language models (Smith et al., 2022).

Despite growing interest in the research community to evaluate gender bias in the classification tasks, most efforts to evaluate bias still do not go beyond gender as binary. Most of the recent work on evaluating gender bias in NLP systems uses variations on Hardt et al.'s work on equal opportunity and equalised odds (Hardt et al., 2016). These measures are group measures and use the gender distributions in the training data rather than the democratic parity measure which insists on equal outcomes for both genders regardless of prevalence or ground truth. Equality of opportunity considers where the predictions are independent of gender but conditional on the ground truth or positive outcome in the training data. This means that the true positive rate of the system should be the same for all genders. An example of this is the $TPR_{gap}$ (Prost et al., 2019), as defined in Equation 1, which measures the differences in the gender-specific true positive rates.

$$TPR_{gap} = \mid TPR_{male} - TPR_{female} \mid \quad (1)$$

The more restrictive equalised odds definition of fairness focuses also on restricting differences in errors across genders. An example is the error rate equality differences such as False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED) (Dixon et al., 2018; Park et al., 2018). These metrics are limited to binary labels and depend on threshold values to separate model output into two classes. To address this limitation, Pinned AUC metrics have been proposed (Dixon et al., 2018), but a follow-up study by the same authors found limitations in this metric (Borkan et al., 2019a). As a result, a new set of threshold-agnostic metrics was proposed by Borkan et al. (2019b) which overcomes the limitations of Pinned AUC metrics related to class imbalance and provides robustness and more nuanced insight into the types of bias present in the model.

These metrics are computed based on the score distributions of both the complete background test data, which consists of every other subgroup except the subgroup under consideration, and the test set subgroup. This means that the performance of the model is evaluated not only on the entire dataset but also on the specific subgroup that is of interest. AUC-based metrics include Subgroup AUC,

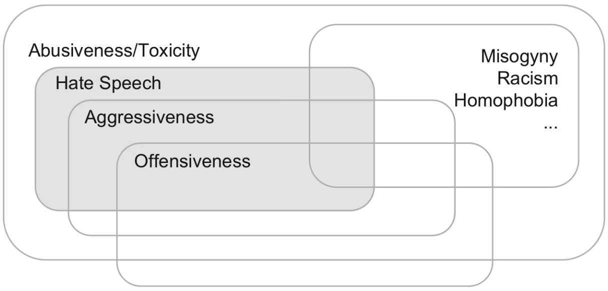

Figure 1: Relations between HS and related concepts (Poletto et al., 2021)

Background Positive Subgroup Negative (BPSN) AUC, and Background Negative Subgroup Positive (BNSP) AUC. Subgroup AUC calculates the measure of separability for a given subgroup, which gives a more accurate understanding of the model's performance in that particular subgroup. While these metrics can be used for measuring different kinds of bias (e.g racism, religious, etc.) across different subgroups, our focus is on gender bias, considering three distinct subgroups: male, female, and gender-neutral. The Background Positive Subgroup Negative AUC (BPSN) metric evaluates the AUC score by considering the positive examples from the background and the negative examples from the subgroup. Lower values in this metric mean more false positives within the subgroup at many thresholds.

On the other hand, the Background Negative Subgroup Positive AUC (BNSP) metric calculates the AUC by considering the negative examples from the background and the positive examples from the subgroup. A low BNSP score presents more false negatives within the subgroup. In other words, low BNSP indicates that more positive examples from the subgroup are mistakenly classified as negative at different thresholds.

The set of metrics also include an Average Equality Gap which measures the difference between true positive rates for each outcome for a subgroup and the background at a specific threshold. This is a generalisation of the TPR_gap in Eqn 1 above across multiple subgroups. Equation 2 shows the AEG for the positive outcome where $D_g^+$ is the positive data for the subgroup $g$, $D^+$ is the positive data in the background i.e. all data except the subgroup, and MWU is the Mann-Whitney U test statistic.

$$PositiveAEG = \frac{1}{2} - \frac{MWU(D_g^+, D^+)}{\mid D_g^+ \mid\mid D^+ \mid} \quad (2)$$

There is an equivalent AEG for the negative outcome for a particular subgroup. The values of AEGs range from -0.5 to 0.5, with an optimal value of 0 indicating no differences between the particular subgroup and the background data.

## 3 Approach and Evaluation

This research aims to explore gender bias in hate speech and offensive language classification, with a specific focus on gender-neutral language. We will accomplish this by analysing commonly used user-generated content datasets, particularly three Twitter datasets for abusive content and offensive language identification. These datasets have been used in prior bias detection studies (Park et al., 2018; Davidson et al., 2019).

Abusive language includes various types including stereotypes, offense, abuse, hate speech, threats, etc (Caselli et al., 2020). The connections among these phenomena based on previous research, have been visually represented in a work by (Poletto et al., 2021) and it is shown in Figure 1. Current approaches for detecting and mitigating harmful language mainly focus on offensive language, abusive language, and hate speech however with varied and inconsistent definitions (Caselli et al., 2020; Waseem et al., 2017). The difference between offensive language, abusive language, and hate speech lies in their specificity. Offensive language is more general, hate speech is more specific, and abusive language falls in between.

The **Hate Speech** dataset (Waseem and Hovy, 2016) is a collection of almost 17K tweets consisting of 3,383 samples of sexist content, 1,972 samples of racist content, and 11,559 neutral samples. The dataset is transformed into a binary classification problem by labeling the sexist and racist samples as the "abusive" class and neutral samples as the "non-abusive" class.

The **Abusive Tweets** dataset is a large-scale

| Dataset | Class | Class% | gender-neutral% | identified gender |  | Size |
|---|---|---|---|---|---|---|
|  |  |  |  | F(%) | M(%) |  |
| Hate Speech | Abusive | 31.4 | 2.0 | 1.0 | 1.0 | 16K |
|  | Non-Abusive | 68.6 | 3.0 | 1.0 | 2.0 |  |
| Abusive Tweets | Abusive | 32.1 | 2.0 | 1.0 | 1.0 | 100K |
|  | Non-Abusive | 67.9 | 3.0 | 1.0 | 2.0 |  |
| Hate Speech/Offensive | Abusive | 50.0 | 4.0 | 2.0 | 2.0 | 8K |
|  | Non-Abusive | 50.0 | 3.0 | 1.0 | 2.0 |  |

Table 1: Class distribution, gender neutral data, gender labeled data percentage, and overall size for each dataset. For the HateSpeech/Offensive dataset, the abusive class has been undersampled due to significant class imbalance.

crowd-sourced dataset, collected by Founta et al. (2018). The size of the dataset is just under 100k tweets and it is annotated with four labels: *hateful*, *abusive*, *spam*, and *none*. By combining the *none* and *spam* instances into a "non-abusive" class, and the hateful and abusive instances into an "abusive" class, we transform the dataset to a binary classification task, similar to the Hate Speech dataset.

The **HateSpeech and Offensive** dataset (Davidson et al., 2017) is a collection of almost 25k tweets. The majority of tweets are considered to be offensive language (77%), almost 17% are labeled as non-offensive and only almost 6% of the tweets are flagged as hate speech samples. By assigning the "abusive" class label to samples exhibiting hate speech and offensive, and the "non-abusive" label to non-offensive samples, we convert the dataset into a binary classification problem.

The HateSpeech & Offensive dataset contains a significant class imbalance, 83% of the dataset is assigned as abusive while only 17% is assigned as a non-abusive class. In order to create a more balanced dataset for experimental purposes, undersampling was performed on the abusive class during the evaluation by randomly selecting a 17% sample of the abusive data leaving a balanced dataset for this work of 8305 instances.

Table 1 shows the characteristics of the data used in the evaluation, including size and class distribution.

| binary | gender-neutral |
|---|---|
| he/she | they |
| him | them |
| her | them,their |
| his | their,theirs |
| hers | theirs |
| himself/herself | themselves |

Table 2: Binary pronouns and neutral alternatives

### 3.1 Gender Neutralising the Training Data

The dataset we used for our analysis lacks gender-neutral language or has a very limited representation of it. To address this issue, we employed the Neutral Rewriter (Vanmassenhove et al., 2021) to generate gender-neutral samples. This model which is a combination of rule-based and neural approaches replaces gender identity terms with their gender-neutral equivalents.

Results from the Neutral Rewriter demonstrate that the model achieves a high level of accuracy, with a word error rate of less than 1%. Table 2 shows the pronouns and their gender-neutral alternatives used by the model. The gender-neutral rewriter also replaces gendered English animate nouns with gender-neutral terms. For instance, "postman" is substituted with "mail carrier," and "fireman" with "firefighter." Similarly, feminine forms of animate nouns such as "actress" are replaced with gender-neutral alternatives like "actor," and "waitress" with "waiter." Additionally, the rewriter replaces generic uses of "man," for instance, "freshman" can be replaced with "first-year student," and "man-made" can be replaced with "human-made. The complete list of mapped nouns to their gender-neutral alternatives could be found in the original paper (Vanmassenhove et al., 2021). As an example, the sentence *she is an actress* would be replaced by *they are an actor*. Label preservation was not checked after gender-neutralising was performed. There may be certain instances that, after gender, may not be considered hateful, particularly for gender stereotyping due to traditional gender roles.

Using the gender-neutral rewriter model, we generated gender-neutral data instances from the original datasets. 60% of the data instances that could be gender-neutralised were replaced with a gender-neutral version and we left the remaining 40% that included gender pronouns/determiners in the dataset, unchanged. It was important not to gender neutralise all instances with specific gender identity terms which could potentially leave the training data without these terms at all.

It should be mentioned that to prevent bias mitigation selected instances were replaced with their gender-neutral version rather than being added to the dataset. This replacement technique aimed to avoid mitigating bias through gender swapping which is a technique to mitigate bias by augmenting the data with additional examples that are gender-swapped (Park et al., 2018).

There is no gender identity feature or label in the datasets used. The approach in Identity Term Sampling (ITS) (Sobhani and Delany, 2023) which is used to identify male and female instances within text data was used to identify gender within the portion of data to be gender neutralised. ITS identifies gender for data instances to allow the measurement of task performance across genders. It identifies gender based on the frequency of occurrence of certain gender identity terms. We are using these proportions identified by ITS as an estimate of gender distribution in the instances changed by the gender-neutralising process. The column labelled *gender-neutral%* in Table 1 shows the gender distribution of the gender-neutralised data instances using this approach.

### 3.2 Generating Test Data

In order to generate synthetic test instances to isolate the gender-neutral, female and male groups we extended an existing template approach that has been used for measuring bias in hate speech (Park et al., 2018) and toxicity classification (Dixon et al., 2018). Tables 4 and 3 present the templates and the identity terms respectively used to generate sets of labeled test data for male/female and gender-neutral scenarios while Table 5 lists the nouns and adjectives used to fill the templates. This template approach generated 3006 synthetic test samples in total, with 1002 samples for each of the male, female, and gender-neutral groups. The samples were evenly distributed across the *abusive* and *non-abusive* classes.

### 3.3 Experimental Approach

For classification, a simpler version of the deep neural network architecture used by (De-Arteaga et al., 2019) is used. Our architecture consists of an embedding layer as an input layer using word2vec embeddings, followed by a bidirectional long short-term memory (BiLSTM) layer to encode the input sentence, a dropout layer, and a linear output layer with cross-entropy to compute the loss which comes after a dense layer with Relu activation.

| Identity Terms |
|---|
| woman/man/adult, girl/boy/kid |
| female/male/person, niece/nephew/nibling |
| sister/brother/sibling, gal/guy/folk |
| mother/father/parent, she/he/they |
| wife/husband/spouse, her/his/them |
| daughter/son/child, aunt/uncle/pibling |
| girlfriend/boyfriend/partner |
| granddaughter/grandson/grandchild |
| grandmother/grandfather/grandparent |

Table 3: Identity terms used to generate sets of test data with an identified gender

| Templates |
|---|
| You are a <*adjective*> <*identity term*> |
| <*verb*> <*identity term*> |
| Being a <*identity term*> is <*adjective*> |
| I am a <*identity term*> |
| I hate <*identity term (plural form)*> |
| <*she/he/they*> is/are <*adjective*> |
| I <*verb*> <*her/him/them*> |

Table 4: Templates used to generate synthetic test data

The datasets were split 80%/20% into stratified training and holdout testing sets. All hyperparameters were tuned on a 20% split of the training data for each dataset. The performance on the holdout test data was averaged across three holdout splits. As the HateSpeech & Offensive dataset is highly imbalanced, the process of randomly undersampling the 17% sample of the abusive data was repeated 5 times to ensure the results were not dependent on a specific subset of the data. The average result across the 5 iterations was then reported.

Overall task performance was measured using accuracy on the synthetic test data. In addition, to gain a better understanding of the model's general performance on the task in all datasets, we measured the average class accuracy (ACA) on a representative proportion of the data. We chose to use ACA as a complementary metric due to the imbalanced class distributions present in the data.

In addition to task performance, we measured gender bias on the synthetic test instances using AUC-based metrics including Subgroup AUC,

| Nouns & Adjectives & Verbs | Target |
|---|---|
| disgusting, filthy, nasty, rotten, horrible, terrible, awful, hate, vile, idiotic, stupid, moronic, dumb, ugly, shitty, fucked, kill, murder, repulsive, destroy | Abusive |
| great, fun, nice, neat, happy, best, good, fantastic, wonderful, lovely, excellent, incredible, friendly, gracious, kind, caring, hug, like, love, respect | Non-Abusive |

Table 5: Nouns, adjectives & verbs used to generate synthetic test data

BPSN, BNSP, and AEG on the positive (*abusive*) and negative (*non-abusive*) classes. Subgroup AUC calculates AUC within a specific subgroup, BPSN measures AUC within the positive background and negative subgroup, and BNSP measures AUC within the negative background and positive subgroup. Additionally, the positive and negative AEG measure the separability of positive (abusive) examples from the subgroup with positive (abusive) examples from the background data and vice-versa. Positive AEG is defined as Equation 2.

## 4 Results and Discussion

Table 6 shows the classification results including overall accuracy and class level accuracy across the three datasets for the classification of hate speech. The table displays the average class accuracy (ACA) for the holdout test data in the last column. This metric indicates how well the model can perform on classifying abusive content, with a higher ACA indicating better performance. The results show a good performance of the model generally - the Abusive Tweets dataset with an ACA of 90%, while the HasteSpeech & offensive dataset has an ACA of 88%, and the Hate Speech dataset has an ACA of 81%. However, looking at the class accuracy column in Table 6 it can be seen that the model performed poorly in classifying abusive content, with less than 50% accuracy across all three datasets. The strong performance on the non-abusive class is contributing to the overall good performance.

Table 6 also shows the performance of the model on the synthetic test dataset. Results show the accuracy on synthetic test data is less than 75% across three datasets, which means the model does not perform as well in classifying the synthetic datasets. This is not surprising as the template sentences used to generate the test data are not fully representative of the actual abusive content in the datasets. However, this synthetic data can still provide valuable insights into potential biases in the models.

Table 7 shows the gender bias results across the three datasets including the AUC-based metrics and the AEG of the positive (abusive) and negative (non-abusive) classes. The subgroup AUC shows a score higher than 0.7 for all datasets across our three gender identity subgroups which indicates that the model is moderately successful in distinguishing between positive and negative examples within female, male, and neutral subgroups.

The high scores on BNSP and BPSN AUC metrics results for the Abusive Tweets dataset show that the model exhibits relatively low bias across all the female and male and neutral subgroups, with high BNSP and Subgroup AUC scores indicating similar performance to the background group.

However, the two hate speech datasets show some level of bias across these figures and it differs between the different subgroups. Interestingly the figures for the male and neutral subgroups on the hate speech datasets are much closer to each other and higher than the female subgroup. Low values in the BPSN and BNSP AUC metrics indicate more bias. So this suggests that the bias for the female subgroup is higher than the male and neutral.

Looking at what these AUC metrics tell us, the BPSN score for females on the hate speech datasets is relatively low with a score of 0.58 in the Hatespeech and 0.78 in the HateSpeech & Offensive dataset. A low BPSN score suggests that the model is more likely to incorrectly classify negative or non-abusive examples from female subgroups as abusive compared to the background groups, which in this case are male and neutral, indicating the model is more likely to predict abuse for the female instances than the male and neutral instances.

On the other hand, the BNSP score for the hate speech datasets is lower for male and neutral subgroups than the female subgroup. Since the BNSP score measures the difference in false negative rates between the subgroup and the background group the low score in the male and neutral subgroups indicates that the model tends to incorrectly classify abusive examples from both the male and neutral subgroups as non-abusive compared to their respective background group. This suggests that

| Dataset | Class | Class Accuracy(%) | Synthetic testset Accuracy(%) | Original testset ACA(%) |
|---|---|---|---|---|
| Hate Speech | Abusive | 37 | 64 | 81 |
| | Non-Abusive | 91 | | |
| Abusive Tweets | Abusive | 47 | 73 | 90 |
| | Non-Abusive | 98.8 | | |
| HateSpeech & Offensive | Abusive | 48 | 71 | 88 |
| | Non-Abusive | 95 | | |

Table 6: Accuracy per class, accuracy on the synthetic test data, and average class accuracy (ACA) for each dataset across three holdout splits.

for these hate speech datasets, male and neutral abusive instances are more likely to be missed than female instances. The BPSN and BNSP for gender-neutral suggest that the model may be more biased against the gender-neutral subgroup compared to the male subgroup, but less biased compared to the female subgroup. Furthermore, the negative values of both the abusive and non-abusive AEG and the Hatespeech and HateSpeech & Offensive datasets suggest that the model is biased towards the female subgroup, as there is a downward shift in scores for this subgroup. This bias is further supported by the low BPSN AUC score, which indicates that the model is more likely to make false positive predictions for the female subgroup compared to the background groups. Specifically, the negative AEG scores indicate that the model is performing worse for the female subgroup than the reference group, which can contribute to the lower AUC score.

Moreover, positive scores for both the abusive and non-abusive AEG for neutral and male suggest that the model might give more weight or importance to certain features in the neutral and male subgroups when classifying positive and negative examples. This means that the model may be more accurate in classifying positive and negative examples from these two subgroups compared to the background group, with the degree of attribute amplification being relatively small. Also, a positive AEG value for the non-abusive class along with a low BNSP indicates that the model is performing better for the male and neutral subgroup for the non-abusive class. Overall, these results suggest that the model may exhibit some bias against the neutral and male subgroups, particularly in terms of false negative rates, but the degree of bias is less severe compared to that shown for the female subgroup.

Looking at the results for Hatespeech and HateSpeech & Offensive datasets we can see that including gender-neutral data in the datasets shows gender bias in the female subgroup, but surprisingly gender-neutral and male results have similar behavior on the bias metrics. There could be several reasons that cause this behavior. Given the novelty and limited usage of gender-neutral terms in many societies, they might appear infrequently in training data. Consequently, Machine Learning models could encounter difficulties in comprehending and generating gender-neutral language. For instance, terms like "nibling/pibling" or "sibling" are uncommon in daily speech, and may limit the model's exposure to gender-neutral language.

Second, gender-neutral forms of specific words, such as "actress" or "waitress," is often associated with the male form, reflecting a common representation found in many datasets. Another possible reason might be that the gender-neutral term "they" is the same as the plural "they" which might confuse the model in distinguishing singular and plural they.

Results show male and gender-neutral subgroups have similar biased behavior according to Table 7. In order to find out what gender direction (male or female) gender-neutral words align better with, we conducted an analysis of gender bias in word2vec embeddings for gender-neutral words. Following the work by Bolukbasi et al. (2016), we projected the gender-neutral words listed in Table 3 onto the gender direction, which is defined as the vector resulting from $\overrightarrow{she}$ -$\overrightarrow{he}$. Table 8 shows the projection result for gender-neutral words with respect to the projection score in the gender direction. Words with negative scores are biased toward the male gender, while words with positive scores are biased toward the female gender. The majority of the words including "child", "spouse", "parent", "grandchild" and "adult" have negative scores, indicating a bias towards the male gender. This suggests that most gender-neutral words are more closely associated with the masculine gender spectrum which aligns with similar behavior on the bias metrics.

| | | AUC | | | AEG | |
|---|---|---|---|---|---|---|
| Dataset | Identity group | SubGroup | BPSN | BNSP | abusive | non-abusive |
| Hate Speech | Female | 0.72 | 0.58 | 0.87 | -0.16 | -0.15 |
| | Male | 0.75 | 0.79 | 0.68 | 0.06 | 0.07 |
| | Neutral | 0.75 | 0.81 | 0.68 | 0.09 | 0.08 |
| Abusive Tweets | Female | 0.99 | 0.98 | 0.99 | -0.03 | -0.07 |
| | Male | 0.99 | 0.99 | 0.98 | -0.05 | -0.06 |
| | Neutral | 0.98 | 0.99 | 0.96 | 0.07 | 0.09 |
| HateSpeech & Offensive | Female | 0.89 | 0.78 | 0.92 | -0.10 | -0.11 |
| | Male | 0.89 | 0.90 | 0.83 | 0.06 | 0.07 |
| | Neutral | 0.84 | 0.88 | 0.83 | 0.09 | 0.07 |

Table 7: Subgroup AUC, Background Positive Subgroup Negative (BPSN), Background Negative Subgroup Positive (BNSP), positive and negative Average Equality Gap (AEG) across female, male, and gender-neutral subgroups

| projection scores | gender-neutral |
|---|---|
| -0.19951084 | child |
| -0.1787668 | parent |
| -0.17748375 | spouse |
| -0.1583447 | grandchild |
| -0.15611757 | adult |
| -0.14471374 | grandparent |
| -0.10091415 | sibling |
| -0.09393246 | folk |
| -0.016291147 | person |
| -0.0070172176 | partner |
| 0.056548793 | they |
| 0.058097813 | them |
| 0.12156674 | kid |

Table 8: Projecting gender-neutral words on the $\overrightarrow{she}$ -$\overrightarrow{he}$ direction in word2vec embedding

## 5 Conclusion

In this paper, we presented an approach for measuring gender bias in a downstream task of identifying abusive or hate speech that considers male, female, and gender-neutral identities. We adjusted existing real-world datasets by gender-neutralising a random subset of instances and extended existing binary template definitions to include identity terms that reflect gender neutrality. Our approach helps address the lack of training data that includes gender-neutral linguistic forms, which is essential for creating more inclusive NLP models and systems by incorporating gender-neutral words through the use of a gender-neutral rewriter. This can lead to more inclusive NLP models and systems. We have evaluated bias towards male, female, and gender-neutral groups and our findings showed that male and gender-neutral groups have similar bias behavior according to the AUC bias metrics, while the female group shows a higher bias compared to the others. This approach can help promote more fair and equitable NLP systems by identifying gender bias in the data.

While our approach aims to address gender bias in abusive and hate speech detection, there are certain limitations to consider. Firstly, the modification of existing datasets by incorporating gender-neutral instances relies on the availability of such data. The scarcity of gender-neutral linguistic forms in real-world datasets can pose a challenge in achieving adequate representation. Secondly, the template-based approach used to generate synthetic test data may not fully capture the nuances and diversity of gender identities, potentially impacting the generalisability of the results. It is important to acknowledge that the concept of non-binary equivalents for binary gender terms is a subject of ongoing debate and individual preference. While a list of suggested non-binary equivalents has been provided in this paper, it is important to recognise that these terms may not be universally agreed upon or applicable to all non-binary individuals.

In future work, we will explore the impact of adjusting datasets to include more gender-neutral identity terms and examine the influence of the dataset size on the results. In addition, a future focus will be on exploring label preservation after gender neutralisation. We will examine the impact of gender-neutralising instances that may be gender stereotypes due to gender roles and consider cases where the resulting text can lose its perceived hatefulness, especially if the assumption is made that the target is a woman/women.

## Acknowledgements

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6183. The author has applied a CC BY public copyright license to any Author Accepted Manuscript version.

## References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

Daniel Borkan, Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019a. Limitations of pinned auc for measuring unintended bias. *arXiv preprint arXiv:1903.02088*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019b. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How conservative are language models? adapting to the introduction of gender-neutral pronouns. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States. Association for Computational Linguistics.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao and Hal Daumé III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle*. *Computational Linguistics*, 47(3):615–661.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. From inclusive language to gender-neutral machine translation. *arXiv preprint arXiv:2301.10075*.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Eric Michael Smith et al. 2022. " I'm sorry to hear that": finding bias in language models with a holistic descriptor dataset. *arXiv preprint arXiv:2205.09209*.

Nasim Sobhani and Sarah Jane Delany. 2023. Identity term sampling for measuring gender bias in training data. In *Artificial Intelligence and Cognitive Science*, pages 226–238, Cham. Springer Nature Switzerland.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *arXiv preprint arXiv:2102.06788*.

Eddie L Ungless, Björn Ross, and Vaishak Belle. 2023. Potential pitfalls with automatic sentiment analysis: The example of queerphobic bias. *Social Science Computer Review*, page 08944393231152946.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. Neutral rewriter: A rule-based and neural approach to automatic rewriting into gender-neutral alternatives. *arXiv preprint arXiv:2109.06105*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.